5

Institutionen för datavetenskap
Linköpings universitet

EXAM
TDDD74 Databaser för Bioinformatik
TDDB77 Databaser och Bioinformatik

May 30, 2015, 14.00-18.00

Help

Dictionary.

Grades

You can get max 33 points. To pass the exam, grade 3, you need 7.5 and 9 points in the practical and theoretical parts of the exam, respectively. For grade 4 and 5, you need 23 and 29 points, respectively.

Questions

Jose M. Peña will visit the room at 16.00.

Instructions

You can answer in Swedish or English. Write clearly. Give relevant and motivated answers only to the questions asked. State the assumptions you make besides those in the questions. None of these additional assumptions should change the spirit of the exercises.

Good luck!

Practical part (15 points)

Question 1. Data modeling with EER diagram (5 p):

1. We want to create a database to store information about a company that has many departments. Each department is of exactly one out of three possible types. Every department has another department as supervisor. However, a supervisor department is always of the first or second type, and never of the third type. Note that not all the departments of first and second types necessarily supervise some other department. Since the supervisor department of a department may change from time to time, we want to store the date when the current supervisor department was appointed as such.

   Draw an EER diagram for the description above. Feel free to add the attributes that you consider necessary. Clearly write down your choices and assumptions in case you find that something in the information above is not clear.

Question 2. SQL (1 + 1 + 2 +1 = 5 p):

Consider the following relational schema

Country(Name, Code, Capital, Area, Population)
Organization(Name, Abbreviation, Established)
IsMember(Organization, Country, Joined)

The attribute *Organization* in the table IsMember is a foreign key reference to *Abbreviation* in the table Organization.
The attribute *Country* in table IsMember is a foreign key reference to *Code* in the table Country.

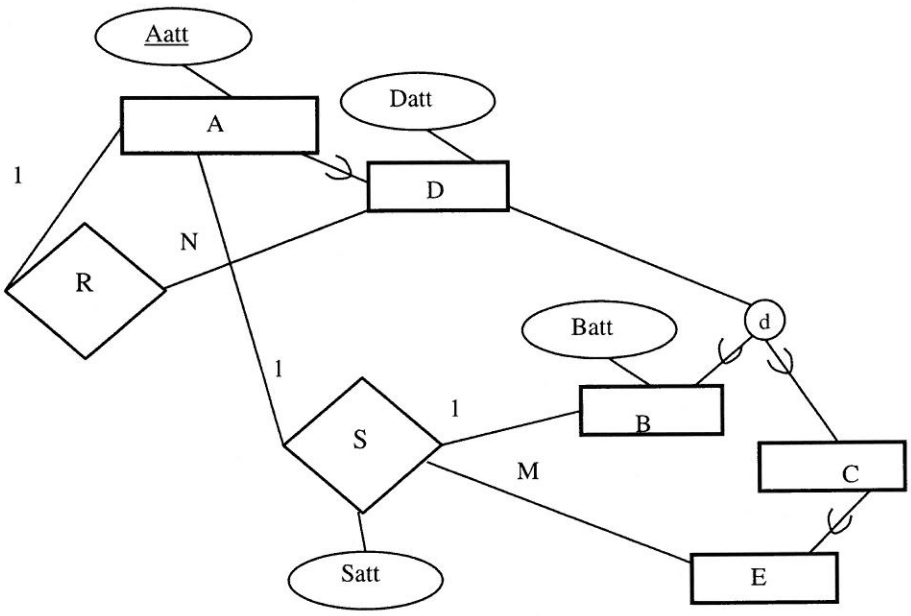
Examples of the tuples from the above relational schema are as follows:

Country(Sweden, SWE, Stockholm, 449964, 9514000)
Organization(European Union, EU, 1952)
IsMember(EU, SWE, 1995-01-01)

1. Show how many members each organization has.
2. Show how many organizations each country joined in the first 10 years since the organization was established.
3. Add a new attribute TotalArea to the table Organization. The attribute represents the sum of the area of the countries belonging to an organization. Fill the attribute with the appropriate values, which you should obtain from the existing tables.
4. Delete all the organizations that were established before 1950.

Question 3. EER diagram and relational schema (5 p):

Translate the EER diagram below into a relational schema with only two tables. Use the algorithm you have seen in the course.



Theoretical part (18 points)

Question 4. Normalization (3 p):

Normalize up to Boyce-Codd normal form (BCNF) the relation R(A, B, C, D, E, F, G) with functional dependencies {A→BCDEFG, BCD→A, BC→EF, D→G, E→F}. Explain your solution step by step. Bear in mind that a relation can have several candidate keys.

Question 5. Data structures (2 + 2 + 1 = 5 p):

We have a file with 1000000 records. Each record is 10 bytes long. The records have two key attributes X and Y. The file is ordered on X. The database uses a block size of B=1000 bytes and unspanning allocation. Each index record is 2 bytes long.

1. Calculate the average (or the maximum, if you prefer) number of block access needed to find a record with a given value for X when using (a) the primary access method and (b) a single level index.
2. Calculate the average (or the maximum, if you prefer) number of block access needed to find a record with a given value for Y when using (a) the primary access method and (b) a single level index.
3. Explain why a single level index performs better than the primary access method in the previous two questions.

Recall that $\log_2 2^x = x$. That is, $\log_2 1 = 0$, $\log_2 2 = 1$, $\log_2 4 = 2$, $\log_2 8 = 3$, $\log_2 16 = 4$, $\log_2 32 = 5$, $\log_2 64 = 6$, $\log_2 128 = 7$, $\log_2 256 = 8$, $\log_2 512 = 9$, $\log_2 1024 = 10$, $\log_2 2048 = 11$, $\log_2 4096 = 12$, $\log_2 8192 = 13$, $\log_2 16384 = 14$, etc.

Question 6. Transactions and concurrency control (1 + 1 + 1 + 1 = 4 p):

Complete the following sentences:

1. Two schedules are conflict equivalent if …
2. A schedule is serializable if …
3. The two-phase locking protocol consists in …
4. The purpose of the two-phase locking protocol is …

Question 7. Database recovery (3 p):

Describe the three recovery methods you have seen in the course. Do not give examples but describe the methods in general terms.

Question 8. Information retrieval (2 + 1 = 3 p):

1. En modell för information retrieval definieras genom 4 komponenter.
   D: hur representeras dokument?
   Q: vilka frågor kan ställas och hur representeras de?

F: vilka dokument ges som svar till en fråga?
R: kan man ge en rankning av svaren och i så fall, hur?

Beskriv *och* ge exempel for D, Q, R, F för booleska modellen.

2. Antag att vi är intresserade av orden: enzym, gene, protein och signal.
Antag att vi har 2 dokument i vår dokumentdatabas.
Antag att dokument 1 innehåller enzym 5 gånger, gen 10 gånger, protein 0 gånger, signal 8 gånger.
Antag att dokument 2 innehåller enzym 0 gånger, gen 0 gånger, protein 7 gånger, signal 1 gång.

Ge dokumentrepresentationerna för dokument 1 och 2 enligt tf-idf modellen.